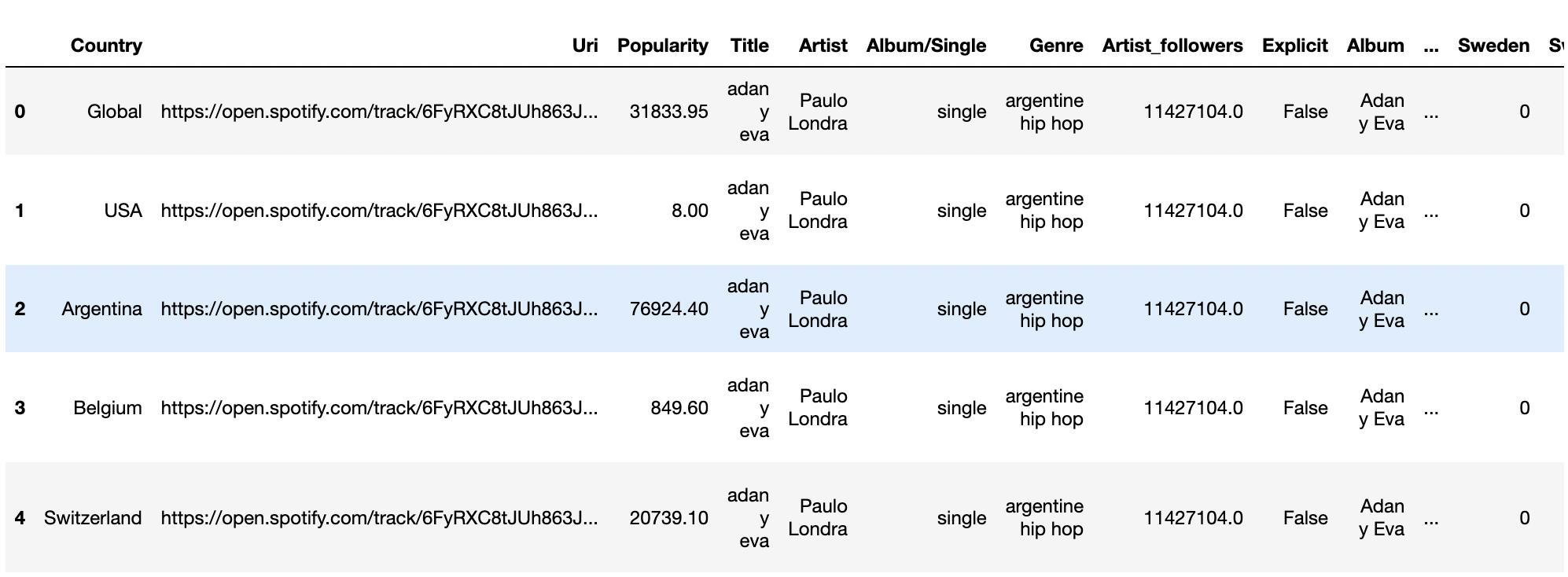
**Project Milestone - Predicting Song Popularity on the Spotify Platform**

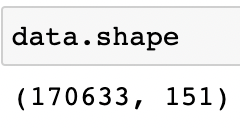
**Exploratory Data Analysis**

We performed some exploratory data analysis on the dataset to understand the features better and to understand the correlations between the features and the target variable which is ‘Popularity’.

The data was loaded into python jupyter notebooks and the pandas package was primarily used for data reading, manipulation and analysis.

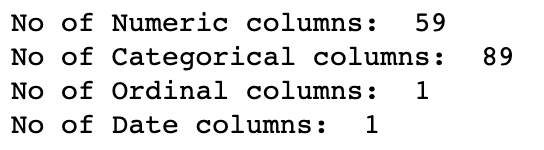
Following is a sample of the data after reading it from the raw .csv file****

The data has a shape of (170633,151). This means we have 170633 unique data points and 150 columns or features and 1 column for the target (Popularity).

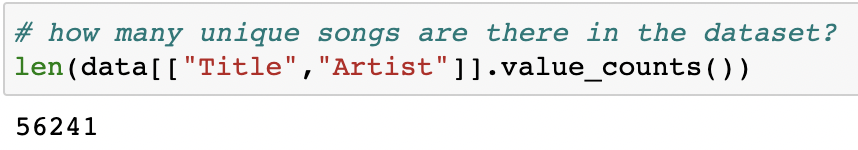
****

We then went through each of the columns and the column description to separate the columns into numeric, categorical, ordinal and date columns.

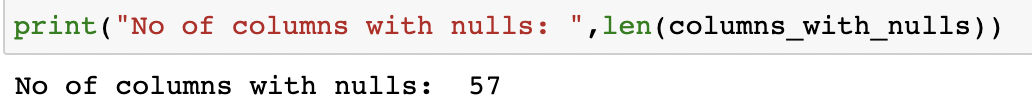
Following is the count of columns in each of the above listed categories -

****

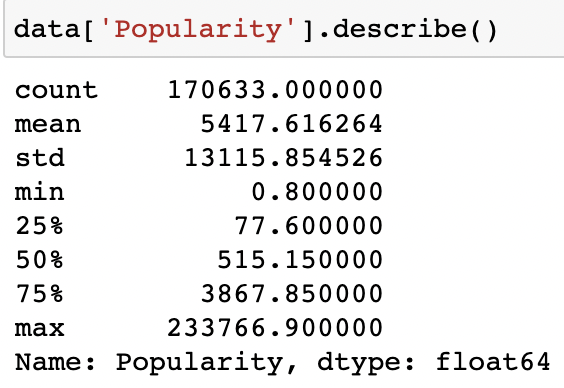
Each data point does not uniquely represent a song in this dataset. As the dataset is that of spotify charts from different countries over 3 years, one song could be showing up multiple times in the data with different popularity each time. We check for unique songs in the dataset for this purpose to help identify features and model better.

****

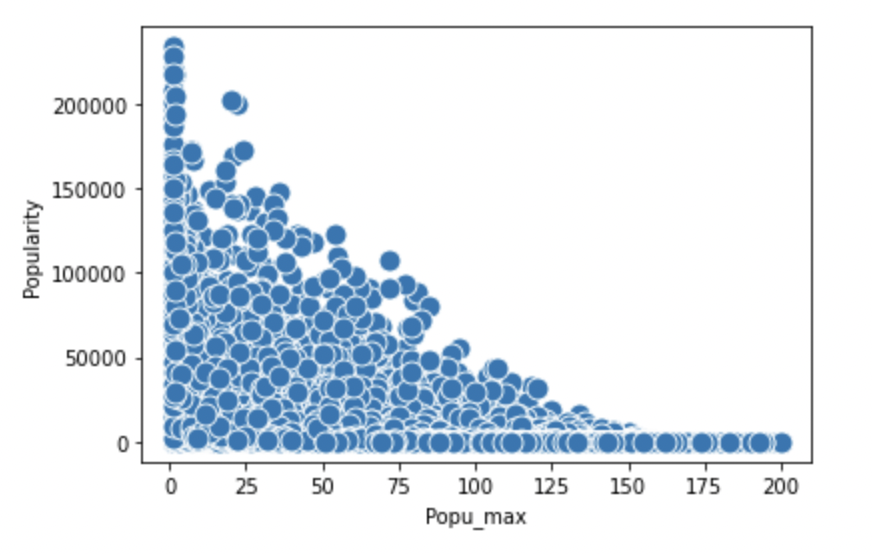
We then checked for columns which have nulls. Although the columns are all not listed here due to space constraints, identifying these columns helps us be aware of the imputations we might have to make before modeling.

****

We then looked at the target variable - ‘Popularity’, and obtained some basic statistics on it to better understand the target. Following is the summary of this basic statistics -

****

One of the variables in the features is ‘Popu\_max’ which is the maximum rank that the song had ever achieved on spotify charts. Ranking on spotify charts is different from popularity and its value lies between 1-200. We wanted to check for correlation between ‘Popu\_max’ and popularity as intuitively it seems like there could be a strong correlation between the two.

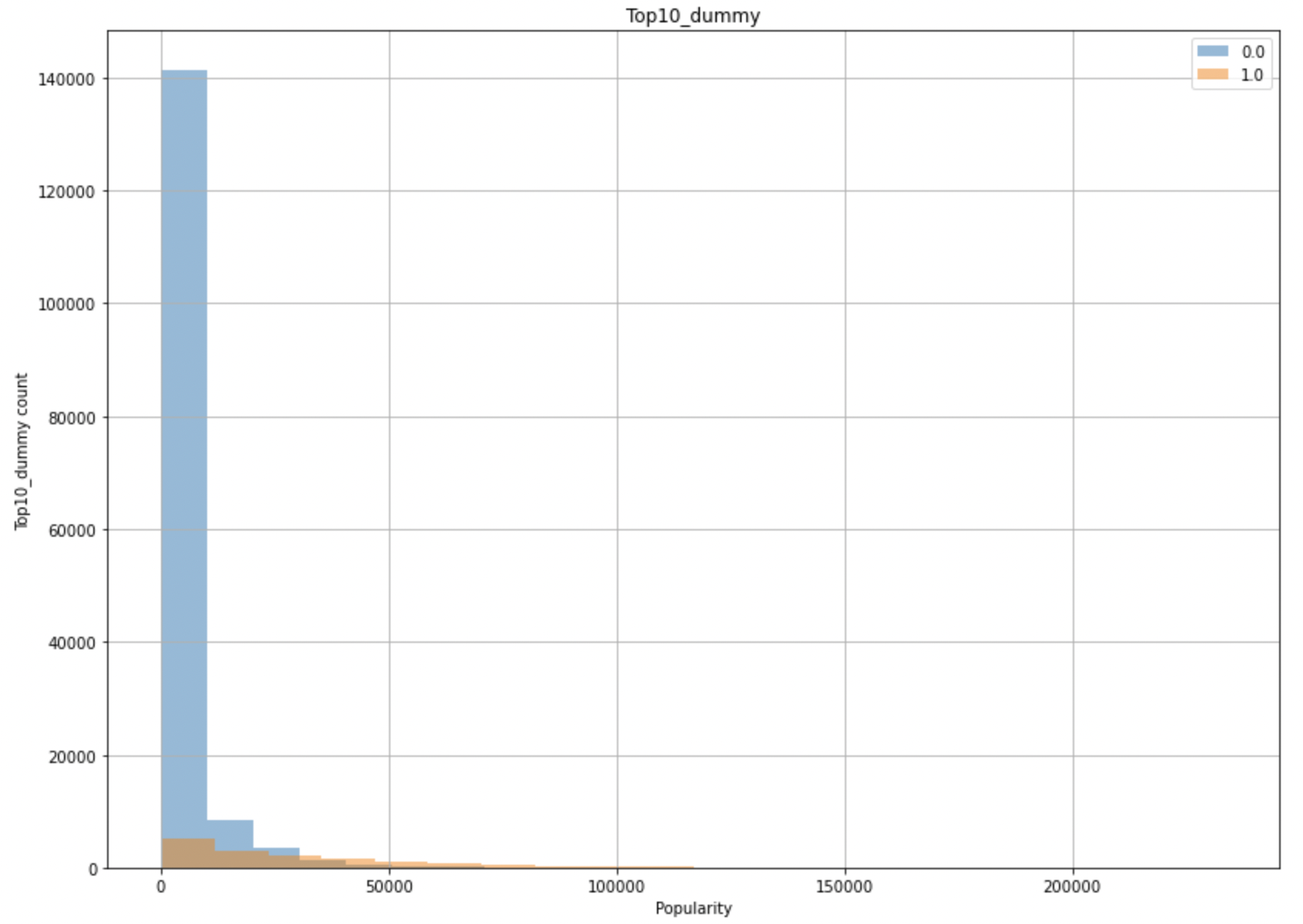
****

However, as the results show above, the correlation between the two as visualized by a scatter plot shows weak relationship with the distribution evenly distributed( almost normally ) and not depicting a strong trend.

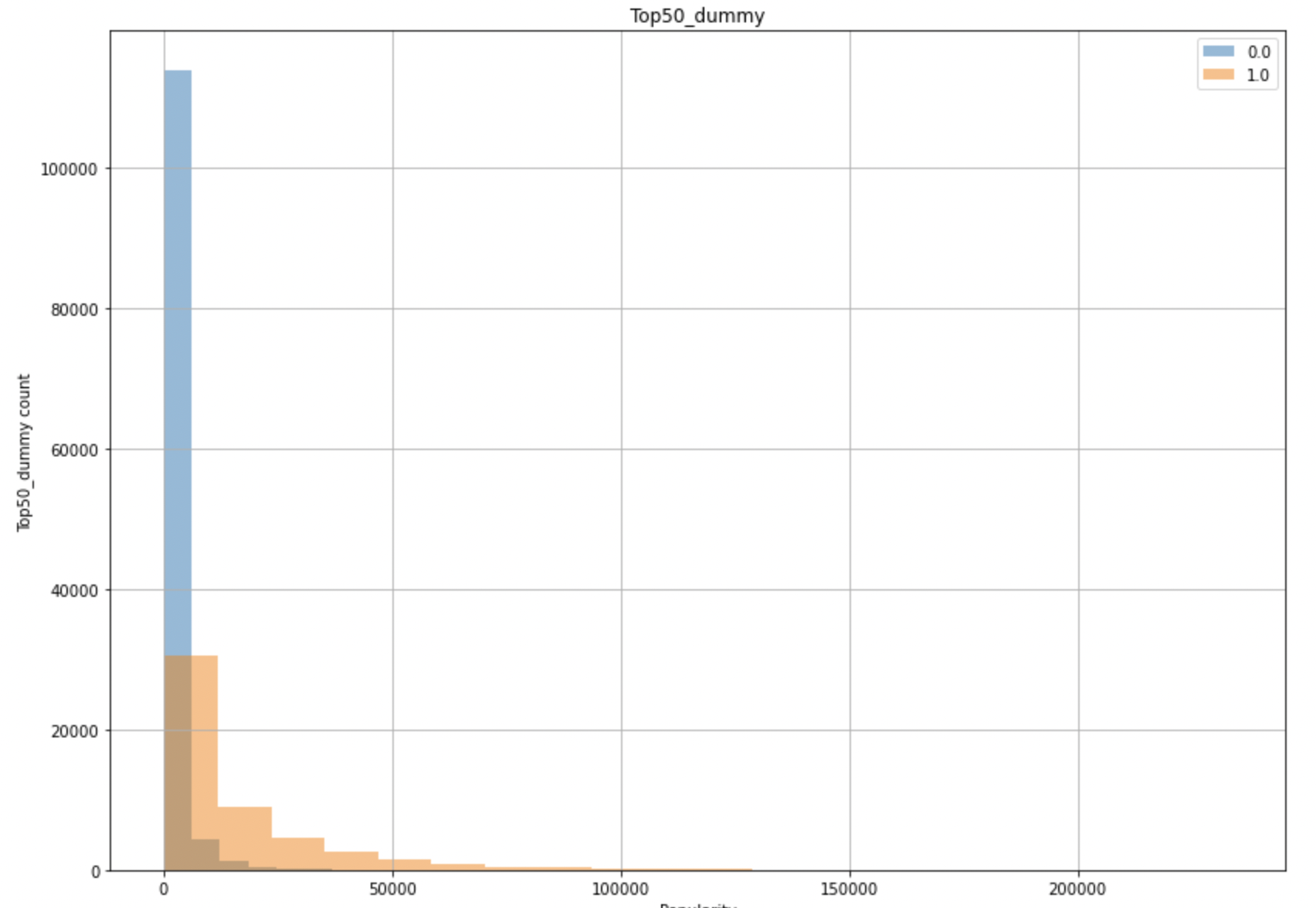
Next we checked the distribution maps of some more variables hypothesized to have strong correlation with popularity for any clear trends in the data.

The ‘top10\_dummy’ feature is a boolean which tells if the song was ever in the top 10 of the charts or not.

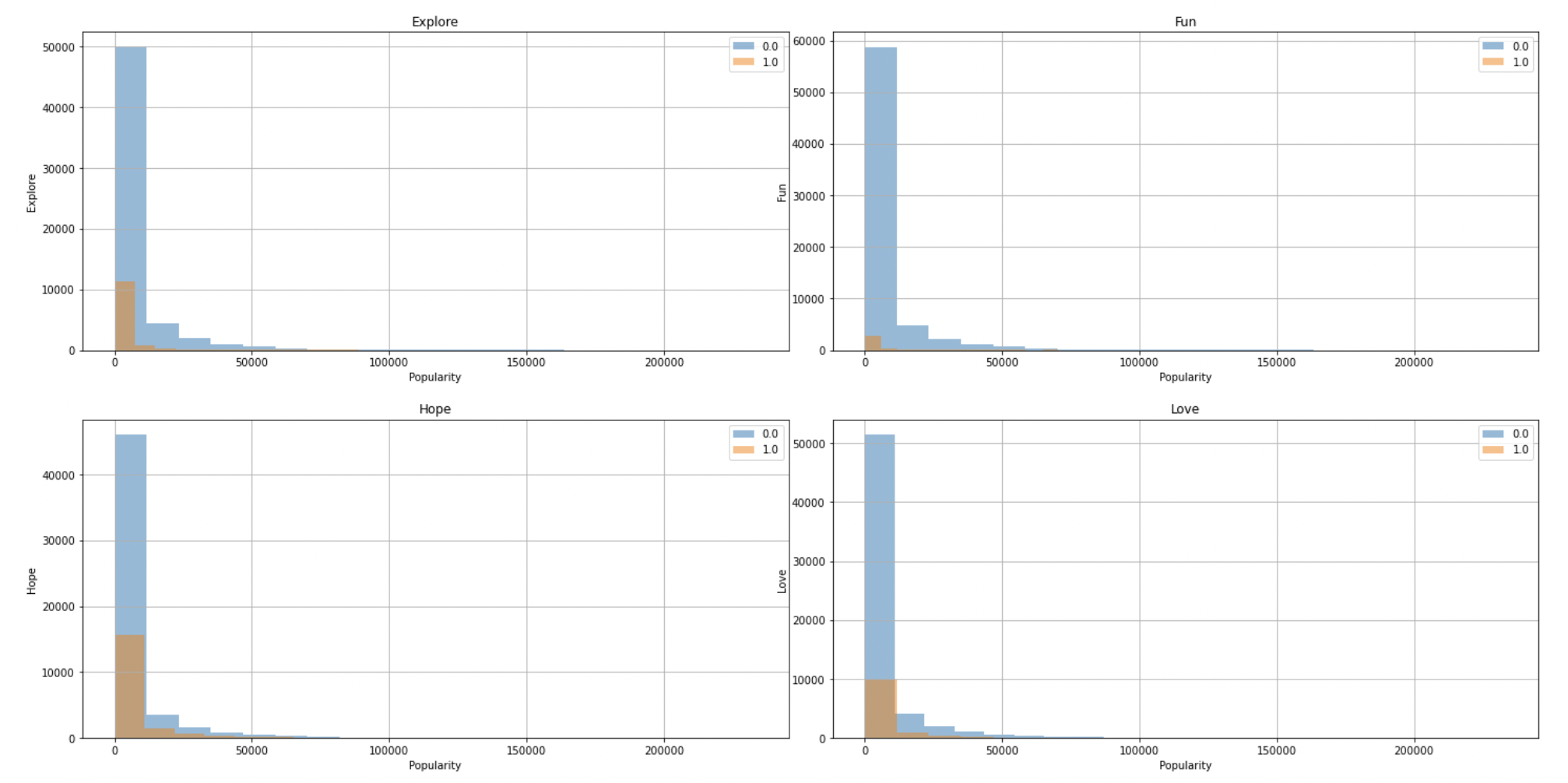
We can see a slightly more right tailed distribution of songs which have been in the top-10 compared to songs that have not.

****

The observation is even more marked in top50\_dummy feature, where one can clearly see that the distribution of songs that have been in top 50 is more right skewed(leaning to higher popularity) compared to distribution of songs that have not been in top 50.

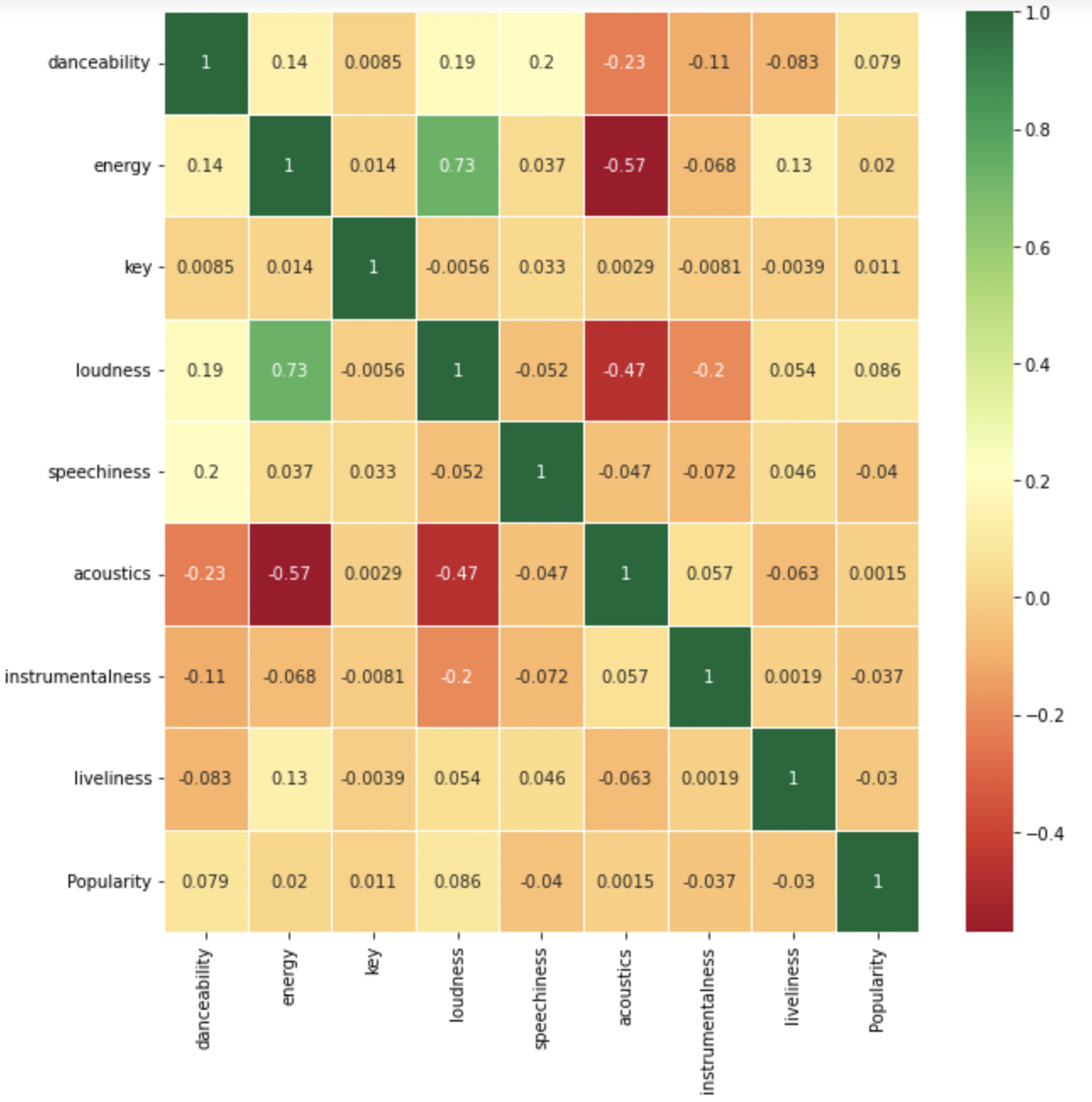
****

We then modeled the mood of the song against popularity using a similar style of visualization. Surprisingly, songs of ‘Hope’ mood seem to have a stronger relationship to popularity over other moods like ‘explore’,’Fun’ ,’Love’ etc.

****

We also visualized the correlation between the song characteristics depicted by a song with each other as well the correlation of each emotion type with popularity.

The heatmap supports common knowledge about songs by showing strong positive correlation between energy-loudness , comparatively(to other correlations)high positive correlation between energy-danceability, strong negative correlation between energy and acoustic ness( usually acoustic songs are melodious and slow).

****

Our preliminary conclusions on the performed EDA is that an apparent and clear correlation between the variables and popularity is lacking in most of the variables, even the ones hypothesized to have a strong correlation. The heatmap also shows very low correlations between song characteristics and popularity. This implies a need for models that can learn complex relations between the variables and the variables and the target in order to predict popularity accurately. More variables will be explored with appropriate visualizations where necessary to understand the features and the data better. The knowledge will be applied to create highly performant models.

**Data Preprocessing and Feature Engineering**

*Overview*

The raw data has 151 columns which comprise factual information about the song (title, artist, release date), popularity score by country (34 countries where Spotify is active + 1 global label), genre information (pop, reggae, etc), musical aspects of the song (tempo, danceability, etc), and sentiment scores generated from the song lyrics (which are **not** included in the dataset). There are also a number of duplicate fields that appear in their original form as sourced from the Spotify API and the one-hot encoded values, e.g. Explicit, explicit\_true, and explicit\_false.

The first steps in preprocessing were identifying the target variables, removing duplicate or unnecessary features, and breaking down any categorical variables into one-hot encoded variables for more effective model training. Once having solidified our initial dataset based on the cleaned and improved features, we then created two datasets that handled null values by either imputing the mean value for the column or converting them to 0s.

*Target Variables*

As we want to not only be able to predict song popularity but also popularity by country, we ideally want to train 35 separate models: one for general popularity and then one each for each of the countries represented in the dataset. This makes our target variables:

| Argentina  Australia  Austria  Belgium  Brazil  Canada  Chile  Colombia  Costa Rica  Denmark  Ecuador  Finland | France  Germany  Global  Indonesia  Ireland  Italy  Malaysia  Mexico  Netherlands  New Zealand  Norway  Peru | Philippines  Poland  Portugal  Singapore  Spain  Sweden  Switzerland  Taiwan  Turkey  UK  USA  Popularity |
| --- | --- | --- |

We will also experiment with both including and excluding the other popularity score variables when training models for specific countries. The reasoning behind this is that when predicting what song will be a hit, an organization may or may not have access to the performance metrics of the song in question in other countries. For these models to be usable in the real world, they would need to account for both scenarios.

*Removing features*

There are a number of duplicate features we will remove that have been doubly accounted for via one-hot encoded variables, such as Explicit, Genre, and Country, as well as [sentiment]\_norm and [sentiment]\_norm2 variables with unclear distinctions. The dataset description explicates the origin of the [sentiment]\_norm2 variables, so we will remove the [sentiment]\_norm variables unless we can find the source of these variables and confirm that they will not be too closely correlated with the other sentiment variables.

In addition, there are a few features that do not seem to have any logical link to the problem statement, such as Released\_after\_2017 and Cluster.

*Adding features*

The raw dataset one-hot encodes only 24 genres, even though there are 1,120 unique genre values. We will replace the existing one-hot encoded values with the full genre set, significantly increasing the number of features.

Performing the above steps leaves us with 1,231 features. As we begin training models and comparing performance metrics we may end up dropping additional sentiment value results if it appears that they have a negative effect on the models’ predictive capabilities, so this is only an initial value.

*Handling null values*

With the feature set created, the next step was to decide how to handle null values. The dataset was designed so that all identifying features of a song, including name, title, artist, genre, and musical facts were present, meaning that we did not have to drop any rows. The null values were by far most prevalent in the sentiment fields, with up to 58% of the dataset missing some or all of the sentiment analysis features. We created two datasets, one with imputed values in place of the nulls and one that converts all null or NaN values to 0. As we train our models and compare performance metrics we will make further observations and decisions on how to handle these features.

**Modeling**

We plan to start the modeling process as the next steps. With the new count of features, we have a lot of data that can be used to train regression models to predict the popularity of a new song based on its musical features. Any song would have to go through a similar process to convert it into features and then we will feed it to the machine learning models to get a prediction of the popularity. We have done some initial research on what regression techniques might work the best. According to our research, Decision Tree, KNN and Random Forest techniques seem to give the best results and we will be using them to train our models. Linear regression does not seem to be working too well. We have also performed these techniques on a smaller subset of this data and they seem to have a similar result.

We will be trying out all the regression techniques and compare the results to select the model with the least root mean squared error to actually predict the popularities for the songs. Based on the final scores that we get, we may also try to use neural network techniques to train models if the performance is not sufficient.